SFOC: A NOVEL MULTI-DIRECTIONAL AND MULTI-SCALE STRUCTURAL DESCRIPTOR FOR MULTIMODAL REMOTE SENSING IMAGE MATCHING

Bai Zhu, Jiachen Zhang, Tengfeng Tang, Yuanxin Ye*

Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610031, China

Commission II, WG II/1

KEY WORDS: Multimodal images, Image matching, SFOC, Fast-NCC$_{\text{SFOC}}$, integral images

ABSTRACT:

Accurate matching of multimodal remote sensing (RS) images (e.g., optical, infrared, LiDAR, SAR, and rasterized maps) is still an ongoing challenge because of nonlinear radiometric differences (NRD) between these images. Considering that structural properties are preserved between multimodal images, this paper proposes a robust matching method based on multi-directional and multi-scale structural features, which consist of two critical steps. Firstly, a novel structural descriptor named the Steerable Filters of first- and second-Order Channels (SFOC) is constructed to address severe NRD, which combines the first- and second-order gradient information by using the steerable filters to depict multi-directional and multi-scale structural features of images. Meanwhile, SFOC is further enhanced by performing the dilated Gaussian convolutions with different dilated rates on it, which can capture multi-level context structural features and improve the ability to resist noise. Then, a fast similarity measure, called Fast Normalized Cross-Correlation (Fast-NCC$_{\text{SFOC}}$), is established to detect correspondences by a template matching scheme, which employs the Fast Fourier Transform (FFT) technique and the integral image to improve the matching efficiency. The performance of the proposed SFOC has been evaluated with many different kinds of multimodal RS images, and experimental results show its superior matching performance compared with the state-of-the-art methods.

1. INTRODUCTION

Image matching is a prerequisite step for remote sensing (RS) image processing and analysis applications, such as image fusion, change detection, and environmental monitoring. The key of RS image matching is to find an evenly distributed and high-precision set of control points (CPs) as much as possible. Generally, RS images can be directly georeferenced by employing the rigorous sensor models or the generic sensor model. However, the georeferencing of RS image is usually biased that caused by the inaccurate measurement of the satellite ephemeris and instrument calibration, which results in the georeferencing having an offset typically ranging from several pixels to dozens of pixels in the image space (Jiang et al., 2015).

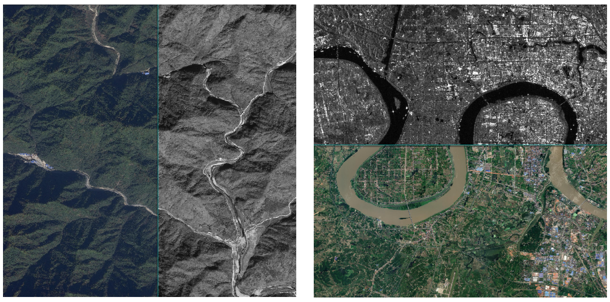


Figure 1. Example of multimodal images with direct georeferencing. (a) Google (Left) and GaoFen-2 Panchromatic (Right) images. (b) Google (Nether) and Sentinel-1 SAR (Upper) images.

Figure 1 exemplarily shows two pairs of multimodal images with direct geo-referencing, and it can be observed that the implementation of georeferencing only can address the obvious global geometric differences. However, there are still significant nonlinear radiometric differences (NRD) between these multimodal images. Moreover, the interference of strong speckle noise is very serious on the SAR image. These challenges make it difficult to detect precise CPs even by visual inspection. Therefore, this paper will focus on developing a robust matching method to resist NRD and noise interference for multimodal RS images.

To date, image matching methods can be commonly classified into three categories with the taxonomy of intensity-based methods (IBM), feature-based methods (FBM), and learning-based methods (LBM). IBM evaluates the similarity of intensity information by using a template matching strategy in the spatial domain or in the frequency domain, which relies on the selection of similarity measures that play a pivotal role in this process. The most common similarity measures consist of four types in the spatial domain: sum of squared differences (SSD), normalized cross-correlation (NCC), mutual information (MI), and phase correlation. Nonetheless, phase correlation, SSD, and NCC are very sensitive to NRD that generally exists in different kinds of multimodal RS images (Ma et al., 2015). Although MI has been testified to be effective for resisting NRD, MI is clumsy and time-consuming because it must compute the joint histogram based on statistical similarity.

FBM differs from IBM to comprise the remarkable features (e.g., point, line, and region features), which evaluates the similarity of these invariant features rather than intensity information to achieve matching. Such methods generally consist of common feature extraction and feature matching, with the most common method to be Scale Invariant Feature Transform (SIFT) (Lowe, 2004) and its variants, such as SAR-SIFT (Dellinger et al., 2014). The above algorithms take advantage of these invariant features to resist geometric distortions, but it is difficult to extract a large

* Corresponding author: Yuanxin Ye, yeyuanxin@swjtu.edu.cn

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-113-2022 | © Author(s) 2022. CC BY 4.0 License.

number of stable features from multimodal images with significant NRD. To tackle these problems, a growing number of valid descriptors have been designed based on structural and shape features. Given the advantages of phase congruency in image perception, numerous phase-congruency-based methods have been proven to improve the performance of multimodal matching (Ye et al., 2017; Li et al., 2019). Although these phase-congruency-based methods have been shown the superiority of the phase congruency in resisting NRD, they required the amplitude and orientation of phase congruency, leading to the complicated calculation and time-consuming processes.

As deep learning has shown superior performance in image matching in the field of computer vision (Dusmanu et al., 2019), LBM has also been introduced into the RS image matching field (Wang et al., 2018b; Zhou et al., 2021). Although current LBMs have achieved remarkable progress, their disadvantages are also quite significant. The main drawback is that LBM usually requires a large amount of training and labeled data, which will greatly affect the practical application of image matching. Due to the number of neural network parameters being huge, the training efficiency is greatly related to the basic configuration of computer infrastructure. LBM's superiority only is brought into play in multimodal image matching based on high-performance computer infrastructures, which is another disadvantage to limit its widespread use.

Recently, many descriptors based on multi-orientated gradient information to depict structural features have also proved to be robust to NRD, among which channel features of orientated gradients (CFOG) (Ye et al., 2019), angle-weighted oriented gradient (AWOG) (Fan et al., 2021), and multi-Scale and multi-Directional Features of odd Gabor (SDFG) (Zhu et al., 2021) are the most representative ones. Moreover, recent studies have shown that many local feature descriptors based on the first-order gradient information, such as SIFT and CFOG, are far from accurate in capturing visual features of human perception. Since the first- and second-order gradients are related to different geometric and structural features of images (Wallis and Georgeson, 2012), the second-order gradients have better performance in describing detailed information than the first-order gradients.

Although the CFOG, AWOG, and SDFG descriptors have been successfully used for multimodal image matching, the construction of gradient channels for CFOG is calculated by simple pixel differences, which are very sensitive to noises. While the horizontal and vertical gradients of AWOG are calculated by the Sobel operator that simply comprises the first-order x-derivative and y-derivative operators. Meanwhile, the multi-scale information is deficient due to both the CFOG and AWOG neglecting the local inter-pixel relationships of images. Although SDFG integrates the multi-scale information for feature description, it is similar to CFOG and AWOG that only make use of the first-order gradients, which results in a lack of local shape attributes in terms of curvature that exploited by the second-order gradients (Huang et al., 2014). Hence, a more discriminative structural feature of the image can be depicted and reinforced when they are used in combination. These observations motivate us to develop a novel descriptor combining the first- and second-order gradient information of images to depict multi-directional and multi-scale structural characteristics.

The main contributions of this paper are composed mainly of two essential components. First, we construct a novel and discriminative descriptor, called the Steerable Filters of first- and second-Order Channels (SFOC), through combining the first-order gradients with the second-order gradients by using the steerable filters, which is utilized to address significant NRD between multimodal images. Then, we establish a fast similarity measure with a template matching strategy, namely Fast Normalized Cross-Correlation (Fast-NCC$_{\text{SFOC}}$), by improving the traditional NCC using the Fast Fourier Transform (FFT) technique and the integral image, which is employed to accelerate the matching process. Therefore, the proposed method with template matching strategy can be regarded as a hybrid method combining IBM and FBM, because it evaluates the SFOC descriptor rather than intensity information to achieve matching.

2. METHODOLOGY

In this section, we will present a novel structural feature descriptor named SFOC based on steerable filters, and it's used to define the fast similarity measure on the basis of NCC. First of all, steerable filters are introduced, consisting of first-order steerable filters and second-order steerable filters. Then, the proposed SFOC descriptor is constructed by utilizing the introduced first- and second-order steerable filter. Finally, the fast-matching similarity measure, namely Fast-NCC$_{\text{SFOC}}$, is developed using the FFT technique and the integral image.

2.1 Introduction of steerable filters

The steerable filters refer to a class of arbitrary orientation filters that can be synthesized into a linear combination of base filters (Freeman and Adelson, 1991). Therefore, the steerable filters can adjust different angles to realize the adaptive control of the filters, with linear, multi-directional, and multi-scale characteristics, so as to provide more details in the image information of direction and edges. The higher-order directional derivatives of the Gaussian function have been proved to be steerable, among which the simplest steerable filter is the first-order Gaussian derivative. The Gaussian function G(x) in two-dimensional space is shown in the following equation:

$$G(x)=\frac{1}{2\pi\sigma^2}e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{1}$$

Where (x, y) are Cartesian coordinates, σ represents the variance of Gaussian function. Let G_n be the nth derivative of the G(x) in the x-direction, and θ represents the rotation of any function concerning the origin. The first-order x Gaussian derivative is expressed as follows:

$$G_{1,\sigma}^{0^\circ}=\frac{\partial G}{\partial x}=(-\frac{1}{2\pi\sigma^4})xe^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{2}$$

If the same function G(x) is rotated 90°, the following equation can be obtained:

$$G_{1,\sigma}^{90^\circ}=\frac{\partial G}{\partial y}=(-\frac{1}{2\pi\sigma^4})ye^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{3}$$

The first-order steerable G_1 filter at arbitrary orientation θ can be synthetized by making use of a linear combination of $G_{1,\sigma}^{0^\circ}$ and $G_{1,\sigma}^{90^\circ}$. Therefore, $G_{1,\sigma}^{0^\circ}$ and $G_{1,\sigma}^{90^\circ}$ are regarded as the basis filters of $G_{1,\sigma}^{\theta}$ filter because all the sets of $G_{1,\sigma}^{\theta}$ can be combined by them.

$$G_{1,\sigma}^{\theta}=\cos(\theta)G_{1,\sigma}^{0^\circ}+\sin(\theta)G_{1,\sigma}^{90^\circ} \tag{4}$$

In addition to the first-order steerable G_1 filter, the second-order steerable G_2 filter is also used in subsequent descriptor construction. Similar to the steerable G_1 filter, the second-order Gaussian steerable filter G_2 is defined as follows:

$$\begin{cases} G_{2,\sigma}^{\theta}=k_1(\theta)G_{2,\sigma}^{0^\circ}+k_2(\theta)G_{2,\sigma}^{60^\circ}+k_3(\theta)G_{2,\sigma}^{120^\circ} \\ G_{2,\sigma}^{0^\circ}=G_{xx}=(-\frac{1}{2\pi\sigma^4})(1-\frac{x^2}{\sigma^2})e^{\frac{-(x^2+y^2)}{2\sigma^2}} \\ G_{2,\sigma}^{90^\circ}=G_{yy}=(-\frac{1}{2\pi\sigma^4})(1-\frac{y^2}{\sigma^2})e^{\frac{-(x^2+y^2)}{2\sigma^2}} \\ G_{xy}=(\frac{xy}{2\pi\sigma^6})e^{\frac{-(x^2+y^2)}{2\sigma^2}},G_2^{60^\circ}=G_{yy}-G_{xy},G_2^{120^\circ}=G_{yy}+G_{xy} \\ k_j(\theta)=\frac{1}{3}[1+2\cos(2(\theta-\theta^j))],\theta_1=0^\circ,\theta_2=60^\circ,\theta_3=120^\circ \end{cases} \tag{5}$$

2.2 Construction of structural feature descriptor

Formally, the construction of SFOC mainly consists of three key components: (1) the construction of the first-order steerable channels with multi-scale strategy, (2) the construction of the second-order steerable channels, and (3) dilated Gaussian convolution and normalization. Figure 2 demonstrates the construction flowchart of the proposed SFOC descriptor and more details of which are specified as follows.

The construction of SFOC is divided into two critical channels: the first-order steerable channels and the second-order steerable channels. Since the convolution operation is a linear operator, thus the first-order steerable channels of an image *I (x, y)* at an arbitrary orientation θ can be computed by convoluting the image with $G_{1,\sigma}^{0^\circ}$ and $G_{1,\sigma}^{90^\circ}$. In the proposed descriptor, the establishment of first-order channels is composed of six directions: $0,\frac{\pi}{6},\frac{2\pi}{6},\frac{3\pi}{6},\frac{4\pi}{6},\frac{5\pi}{6}$. Meanwhile, the multi-scale strategy with different Gaussian standard deviations (STD) is embedded to further reinforce the descriptive completeness of local structural features with the purpose of increasing the discrimination. The specific calculation process is as follows:

$$\begin{cases} S_{1,\sigma}^{0^\circ}=G_{1,\sigma}^{0^\circ}*I(x,y) \\ S_{1,\sigma}^{90^\circ}=G_{1,\sigma}^{90^\circ}*I(x,y) \\ S_{1,\sigma}^{\theta}=\cos(\theta)S_{1,\sigma}^{0^\circ}+\sin(\theta)S_{1,\sigma}^{90^\circ} \end{cases} \tag{6}$$

Where σ represents the Gaussian standard deviation, and $*$ denotes convolution operation.

Furthermore, in order to enhance the detailed information of images, thus the second-order gradients based on the three basic filters (i.e., $G_{2,\sigma}^{0^\circ}$, $G_{2,\sigma}^{60^\circ}$ and $G_{2,\sigma}^{120^\circ}$) are applied in the construction process of the second-order channels. Similarly, the second-order steerable channels of the image *I (x, y)* at an arbitrary orientation θ can be computed by convoluting the image with $G_{2,\sigma}^{0^\circ}$, $G_{2,\sigma}^{60^\circ}$ and $G_{2,\sigma}^{120^\circ}$, which is expressed as Eq. (7).

$$\begin{cases} S_{2,\sigma}^{0^\circ}=G_{2,\sigma}^{0^\circ}*I(x,y \\ S_{2,\sigma}^{60^\circ}=G_{2,\sigma}^{60^\circ}*I(x,y) \\ S_{2,\sigma}^{120^\circ}=G_{2,\sigma}^{120^\circ}*I(x,y) \\ S_{2,\sigma}^{\theta}=\cos^2(\theta)S_{2,\sigma}^{0^\circ}+\sin^2(\theta)S_{2,\sigma}^{60^\circ}-2\sin(\theta)\cos(\theta)S_{2,\sigma}^{120^\circ} \end{cases} \tag{7}$$

Once the synthetical first- and second-order steerable channels are constructed, the specified direction features at different scales are summed to obtain as much useful information as possible in each direction. Subsequently, these synthetical steerable channels in specified directions are convoluted by three parallel Dilated (or Atrous) Gaussian kernels, then the three parallel convolutional results are combined through one summation operation, which is designed to integrate a wealth of local inter-pixel information of images. The dilated Gaussian convolution with different dilated rates by inserting "holes" in the convolution kernels to expand its receptive field, which is inspired by the recent deep convolutional neural networks (Chen et al., 2017). In addition, the dilation rates *r* are set to [1, 2, 3] for avoiding the inherent "gridding" problem that exists in the current dilated

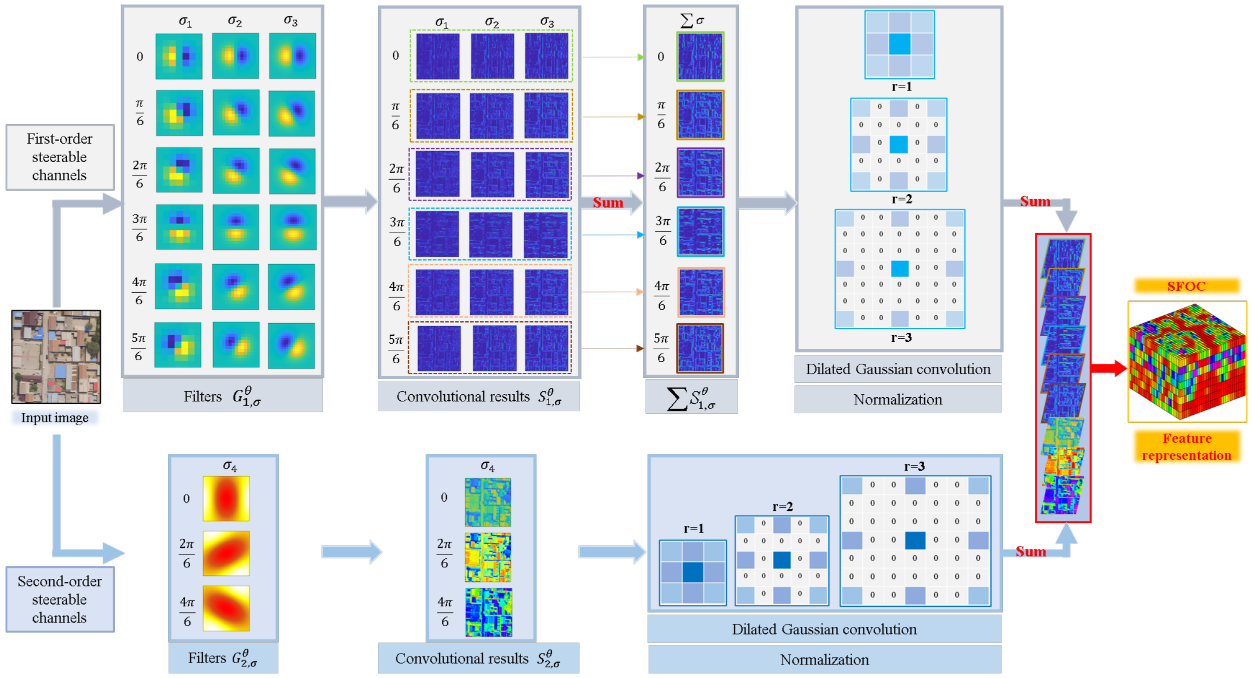

First-order steerable channels
Input image
Second-order steerable channels
σ_1
σ_2
σ_3
0
$\frac{\pi}{6}$
$\frac{2\pi}{6}$
$\frac{3\pi}{6}$
$\frac{4\pi}{6}$
$\frac{5\pi}{6}$
Filters $G_{1,\sigma}^{\theta}$
Convolutional results $S_{1,\sigma}^{\theta}$
Sum
$\sum\sigma$
$\sum S_{1,\sigma}^{\theta}$
r=1
r=2
r=3
Dilated Gaussian convolution
Normalization
Sum
SFOC
Feature representation
σ_4
Filters $G_{2,\sigma}^{\theta}$
Convolutional results $S_{2,\sigma}^{\theta}$
Dilated Gaussian convolution
Normalization
Sum

Figure 2. Construction flowchart of the proposed SFOC descriptor.

convolution framework (Wang et al., 2018a). By this means, the multi-level context structural features of the synthetical first- and second-order steerable channels can be captured by utilizing dilated Gaussian weighting without increasing the computational complexity, and play a role in smoothing noise as well.

Figure 3 clearly illustrates the advantages of utilizing the dilated Gaussian convolution for the construction of SFOC. Four different types of heatmaps concerning different features are acquired by performing template matching. It is obvious that the heatmap of the original image pairs is the messiest, and the heatmap of the SFOC features without Gaussian convolution has several peaks but the peak is not distinct, because it's greatly affected by significant noise. In contrast, Gaussian convolution can effectively resist the interference of noise and make the peak more discriminative (see Figure 3 (e) and (f)). Furthermore, the dilated Gaussian convolution can not only smooth the noise, but also integrate the multi-level context structural features by the dilated Gaussian weighting. This is the reason why the heatmap of the SFOC features with parallel Dilated Gaussian convolution presents a smoother and more discriminative peak than the general Gaussian convolution, which indicates the matching robustness of SFOC with parallel Dilated Gaussian convolution may be superior.

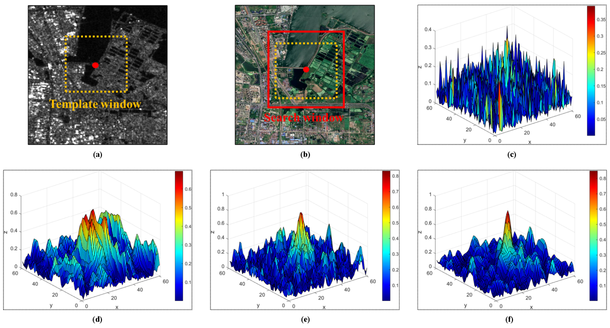


Figure 3. Illustration of the constructed descriptor utilizing different Gaussian convolution strategies. (a) SAR image. (b) optical image. (c) Heatmap of the original image pairs. (d) Heatmap of the SFOC features without Gaussian convolution. (e) Heatmap of the SFOC features with the general Gaussian convolution. (f) Heatmap of the SFOC features with parallel Dilated Gaussian convolution.

In particular, compared with the synthetical first-order steerable channels, a larger σ is used for the dilated Gaussian smoothing for the synthetical second-order steerable channels. This is because the second-order gradient describes more image detail but is accompanied by an increase in noises. Subsequently, the first- and second-order steerable channels are normalized respectively, then the final feature representation of SFOC is obtained by stacking them.

2.3 Establishment of fast similarity measure

The traditional normalized cross-correlation (NCC) is widely applied to determine corresponding CPs between the given image pairs with overlapping regions by evaluating the intensity similarity. However, it is often used only for CP detection of single-modal images and is often unable to keep the same performance for multimodal image matching. As mentioned above, the SFOC descriptor can capture the structural features of images, which effectively resists NRD between multimodal images. Accordingly, it makes sense to establish a novel similarity measure that combines NCC with the SFOC descriptor. SFOC is a 3D descriptor with a large amount of data, as well as the NCC also has the disadvantage of large calculation amount and high computational complexity. Hence, in order to maintain the matching accuracy and improve the computational efficiency, a fast-matching similarity measure is designed based on NCC and SFOC, it's expressed as Fast-NCC_{SFOC}. The proposed Fast-NCC_{SFOC} can be reformulated with more detail as follows in this subsection.

First of all, the SFOC descriptor is used to calculate the structural features in the template image and the search image, which are denoted by T and S, respectively. Their normalized correlation value NCC_{SFOC} (S, T) represents the similarity of the template window $T\,(i, j, z)$ and the search window $S\,(x, y, z)$ at the location (x, y), which is defined as.

$$NCC_{SFOC}(S,T)=\frac{\sum_{h=1}^{z}\sum_{i=1}^{m}\sum_{j=1}^{n}[S(x+i,y+j,z)-\overline{S}][T(i,j,z)-\overline{T}]}{\sqrt{\sum_{h=1}^{z}\sum_{i=1}^{m}\sum_{j=1}^{n}[S(x+i,y+j,z)-\overline{S}]^2\sum_{h=1}^{z}\sum_{i=1}^{m}\sum_{j=1}^{n}[T(i,j,z)-\overline{T}]^2}} \tag{8}$$

Where z presents the dimension of the SFOC descriptor. $S\,(i, j, z)$ and $T\,(i, j, z)$ are the feature value of the search window and the template window at the position (i, j, z), respectively. The sizes of the template and search window are m×n×z pixels and M×N×z pixels, respectively. $\overline{T}$ represents the average feature value of the template image, and $\overline{S}$ represents the average feature value of the search image S under the current template image T.

The reason for the high computational complexity of traditional correlation matching is that NCC is completely recalculated for any search position (x, y), while the internal relation of the NCC of adjacent search points is ignored. In order to reduce the computational complexity, an equivalent transformation is performed on Eq. (8), as follows:

$$NCC_{SFOC}(S,T)=\frac{R_{ST}(x,y,z)-R_S(x,y,z)R_T(i,j,z)/mnz}{\sqrt{[R_{SS}(x,y,z)-R_S^2(x,y,z)/mnz]}\sqrt{[R_{TT}(i,j,z)-R_T^2(i,j,z)/mnz]}} \tag{9}$$

There are only three items related to (x, y, z) are included in the above formula, which is respectively denoted as:

$$R_{ST}(x,y,z)=\sum_{h=1}^{z}\sum_{i=1}^{m}\sum_{j=1}^{n}S(x+i,y+j,z)T(i,j,z) \tag{10}$$

$$\begin{cases} R_S(x,y,z)=\sum_{h=1}^{z}\sum_{i=1}^{m}\sum_{j=1}^{n}S(x+i,y+j,z) \\ R_{SS}(x,y,z)=\sum_{h=1}^{z}\sum_{i=1}^{m}\sum_{j=1}^{n}S^2(x+i,y+j,z) \end{cases} \tag{11}$$

$$\begin{cases} R_T(i,j,z)=\sum_{h=1}^{z}\sum_{i=1}^{m}\sum_{j=1}^{n}T(i,j,z) \\ R_{TT}(i,j,z)=\sum_{h=1}^{z}\sum_{i=1}^{m}\sum_{j=1}^{n}T^2(i,j,z) \end{cases} \tag{12}$$

It should be noticed that the first term $R_{ST}(x,y,z)$ in the numerator is convolution operation, and the convolution in the spatial domain is equivalent to the dot production operation in the frequency domain. Therefore, it can be converted to the frequency domain, and the FFT technique is used to improve computational efficiency. Accordingly, the new expression of the term is equivalent to the following form:

$$R_{ST}(x,y,z)=\int^{-1}[\int(S)\int^{*}(T)] \tag{13}$$

Where $\int$ is the signal of the Fourier transform, $\int^{*}$ represents the conjugate complex operation of the transformed result, and $\int^{-1}$ denotes the inverse FFT (i.e., IFFT).

Additionally, the terms in the denominator and the other terms in the numerator of Eq. (9) require a lot of multiplications and additions. When the template is sliding, the sum of the squares and correlation values are recalculated, which results in computation time increased enormously. It can be seen that these terms, R_S and R_{SS}, fit the definition of the integral image (Viola and Jones, 2001). While the other two terms, R_T and R_{TT}, are only related to the template image, which results in their values being fixed. Therefore, the integral image is used to replace the original summation process with three simple addition and subtraction operations, which can effectively reduce the computational complexity of the original algorithm to calculate NCC, and improve the running time.

As a result, these terms, R_S and R_{SS}, can be efficiently calculated utilizing the integral image. Since the integral process only involves a limited number of additional operations, the complexity of the algorithm is mainly determined by FFT and IFFT in Eq.(13). Typical FFT and IFFT require about $2MNz\log_2(MN)$ times of multiplication, and the Eq. (13) requires to calculate FFT and IFFT once in total. Accordingly, the total number of multiplications required by the proposed Fast-NCC_{SFOC} is as follows.

$$T_1 \approx 4MNz\log_2(MN) \tag{14}$$

With regard to the template matching strategy, the Eq. (8) is directly used to calculate NCC at each sliding position, and the calculation amount mainly depends on the dominant times of multiplication operation. For any search position, the Eq. (8) is used to calculate NCC for about three times of multiplication, and a total of $(M\text{-}m\text{+}1) \times (N\text{-}n\text{+}1)$ slidable positions need to be calculated for traversal search in the search window space. Thus, the number of multiplication operations required for NCC matching is:

$$T_2 = 3mnz(M-m+1)(N-n+1) \tag{15}$$

From the Eqs. (14) and (15) , we can see that the computational complexity of the proposed Fast-NCC$_{\text{SFOC}}$ is independent of the template size, whereas the computational complexity of NCC is approximately proportional to the product of the template size and the search size, especially when m and n are small relative to M and N. The ratio of the computational complexity of the two similarity measures is:

$$T = \frac{T_1}{T_2} \approx \frac{4MNz\log_2(MN)}{3mnz(M-m+1)(N-n+1)} \tag{16}$$

To facilitate the illustration of the computational advantage of Fast-NCC$_{\text{SFOC}}$, we assume that M=N, m=n, and M=2m. The curve of T changing with m is shown in Figure 4. As the template and search sizes increase, the ratio of the computational complexity between Fast-NCC$_{\text{SFOC}}$ and NCC decreases rapidly, that is, the larger the template and search sizes are, the greater the computational advantage of Fast-NCC$_{\text{SFOC}}$ is. Taking a template window m=100 as an example accompanied by a search window M=200, Fast-NCC$_{\text{SFOC}}$ takes about 0.799% of the time required by NCC, which greatly improves the computational efficiency.

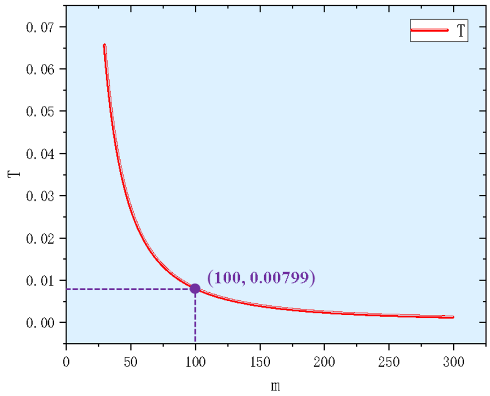


Figure 4. Graph of the variation of T with m.

3. EXPERIMENTS

In this section, the performance of the proposed SFOC was experimentally evaluated with different types of multimodal RS datasets (e.g., optical, infrared, LiDAR, SAR, and rasterized maps). Firstly, the experimental settings were presented, which include the detailed information of test datasets, the evaluation criteria, the implementation details, and the parameters predefined. Then, SFOC was compared with the five state-of-the-art methods for verifying its effectiveness, including MI, matching by tone mapping (MTM) (Hel-Or et al., 2013), phase congruency structural descriptor (PCSD) (Fan et al., 2018), CFOG, and SDFG. Finally, we analysed the robustness of SFOC against Gaussian white noise and speckle noise.

3.1 Experimental settings

Eight cases of multimodal image pairs with significant NRD were employed to evaluate the performance of SFOC. The detailed information of these cases is given in Table 1, and these image pairs of each case are displayed in Figure 6. In addition, the two images of each case have been pre-registered with the same ground sample distance (GSD) to remove obvious rotation and scale differences.

Category		Image source	Size and GSD	Data
Optical -to- Infrared	1	Daedalus optical Daedalus infrared	512×512, 0.5m 512×512, 0.5m	04/2000 04/2000
	2	QuickBird visible QuickBird infrared	1028×1137, 2.4m 1028×1137, 2.4m	05/2006 05/2006
LiDAR -to- Optical	3	LiDAR intensity WorldView-2	600×600, 2m 600×600, 2m	10/2010 10/2011
	4	LiDAR depth WorldView-2	524×524, 2.5m 524×524, 2.5m	10/2010 10/2011
Optical -to- SAR	5	Sentinel-2 optical Sentinel-1 SAR	1501×1501, 10m 1501×1501, 10m	09/2018 10/2018
	6	Google Earth TerraSAR-X	628×618, 3m 628×618, 3m	03/2009 01/2008
Optical -to- Map	7	Google Maps Google Maps	700×700, 0.5m 700×700, 0.5m	unknow unknow
	8	Google Maps Google Maps	621×614, 1.5m 621×614, 1.5m	unknow unknow

Table 1. Detailed information of all test cases

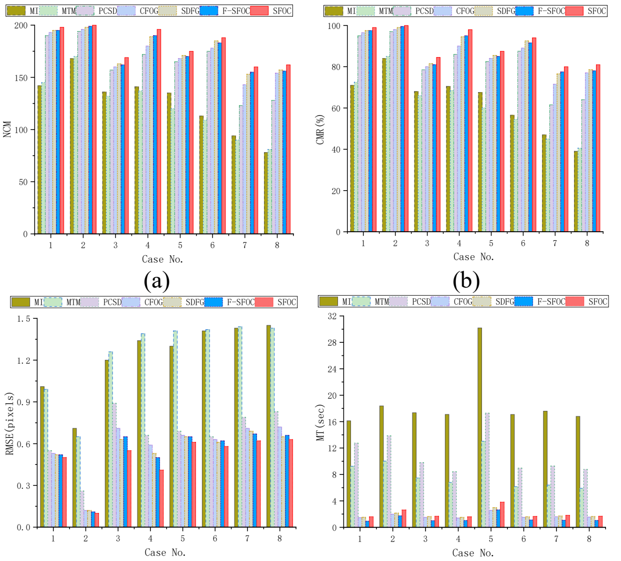
In the experiments, the block-based FAST operator was first employed to extract 200 uniformly distributed IPs from the reference image. Then, the CP detection was performed using different methods with the same template size of 80 × 80 pixels. Furthermore, four criteria were used to quantitatively evaluate the matching performance in terms of the number of correct matches (NCM), the correct matching ratio (CMR), the root-mean-square errors (RMSE), and the matching time (MT). The correct match was determined by manually selecting 50 evenly distributed CPs to estimate the projective model for the image pairs of each case. The projective model was used to calculate the location errors of the matches obtained by different methods, and the match within positioning errors of 1.5 pixels was defined as the correct CP. CMR was defined as CMR = NCM / total matches, where total matches refer to all matched CPs, including the outliers with large errors.

To make a fair comparison, MI was calculated using a histogram with 32 bins, as this is usually accompanied by an optimal matching performance (Ye et al., 2019). And the parameters of the other comparative methods (i.e., MTM, PCSD, CFOG, and SDFG) used the best parameters recommended in their related papers. All experiments were performed using a personal computer (PC) with the configuration of Inter (R) Core (TM) CPU i7-10750H 2.6GHz and 16GB RAM.

3.2 Comparison and analysis of matching performance

In this section, the performance of the proposed method was quantitatively and qualitatively evaluated. Moreover, to evaluate the effectiveness of the second-order gradient in the generation process of SFOC, the SFOC descriptor was degraded by only using the first-order steerable channels without the second-order steerable channels. The degraded SFOC descriptor was represented by F-SFOC, and it was also used for matching performance comparison with other methods.

The seven different methods, i.e., MI, MTM, PCSD, CFOG, SDFG, F-SFOC, and SFOC, were applied to eight multimodal image cases (Table 1) for the comparison of matching performance. Figure 5 depicts the comparison results of all the evaluation criteria (i.e., NCM, CMR, RMSE, and MT) for the different methods on each multimodal image pair. It is obvious that SFOC outperformed the other methods for the above four criteria in all test cases, which effectively demonstrates the superiority and robustness of the proposed SFOC.



(a) (b)

(c) (d)

Figure 5. Performance comparison of different methods on the eight multimodal image cases with the template size of 80 × 80 pixels. (a) NCM. (b) CMR. (c) RMSE. (d) MT.

Among the six methods used for comparison, the worst matching performance was found in the MI and MTM. MI and MTM had comparable matching performance, but MTM performed slightly better than MI on two Optical-to-Infrared cases, while MI performed better than MTM on cases 3-8. This may be related to the fact that MTM only utilizes a piecewise linear function to fit the intensity changes widely existing in the multimodal images. However, the intensity relationship between optical and SAR (or LiDAR) images is too complex to be fitted by MTM, which results in its performance degradation. Although the performance of MI was slightly better than that of MTM, it was the most time-consuming among all the methods because it requires calculating a large number of joint probability histograms.

From the comparison results in Figure 5, we can also observe that PCSD, CFOG, and SDFG performed significantly better than MI and MTM, while SDFG had slightly better performance compared with CFOG and PCSD. The main reason is that PCSD is constructed by using the multi-scale phase congruency structural features, and CFOG is built making use of the dense channel features of orientated gradients, which is more robust to NRD than MI and MTM. When comparing PSCD with CFOG, its performance was slightly worse than that of CFOG. The reason for that is the PCSD may lose some detailed structural information because it employed the strategy of the phase congruency order-based region division for descriptor construction, As for SDFG, since it further increasingly adopted the multi-scale strategy on the basis of multi-direction using odd Gabor functions, its matching performance was more robust than CFOG, but the matching process was more time-consuming. In addition, the construction of PCSD relies on multi-scale phase congruency features, which results in it being time-consuming. Therefore, PCSD and MTM were the most time-consuming apart from MI in all the compared methods.

For our degraded descriptor (i.e., F-SFOC), its matching performance was comparable to SDFG, and it yielded better results than CFOG on the criterion of RMSE, especially in the LiDAR-to-Optical and Optical-to-SAR cases. This phenomenon illustrates that the first-order Gaussian steerable filters and the dilated Gaussian convolution are effective to construct the descriptor. While the matching performance of F-SFOC was obviously lower than SFOC, which verified the feasibility and effectiveness of adding the second-order gradient in the generation of SFOC. In this way, the robustness and discriminability of SFOC can be effectively increased. As far as the MT, F-SFOC was slightly faster than CFOG, because it only took advantage of the first-order steerable channels without the second-order steerable channels resulting in a smaller dimensionality of its features than that of CFOG. Whereas SFOC required slightly more time-consuming than CFOG and SDFG, this is related to the multi-scale strategy with different Gaussian STD and the dilated Gaussian convolution with different dilated rates were embedded in the generation process of SFOC. Hence, considering the improvement of the matching performance for SFOC, it is acceptable to sacrifice a little running time.

Moreover, qualitative evaluation was performed by displaying the correct matched CPs for the visual inspection. As shown in Figure 6, these CPs were established by SFOC on the image pairs of each case with a template size of 80 × 80 pixels. it is obvious that these obtained CPs on the image pairs of each case were evenly distributed, and the location accuracy of these CPs was reliable despite significant NRD and noise between these multimodal image pairs.

3.3 Comparison and analysis of noise sensitivity

In this section, the anti-noise performance of the above-mentioned methods was evaluated and analyzed by adding

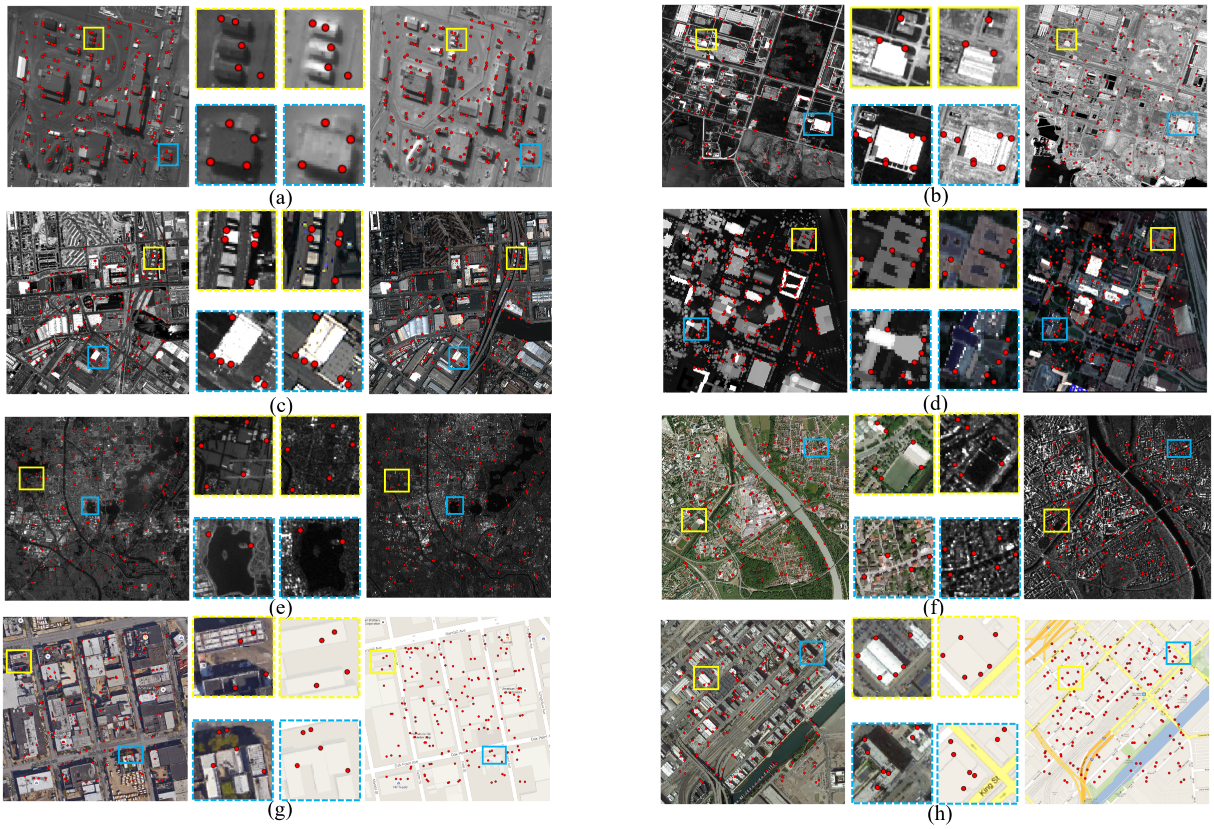
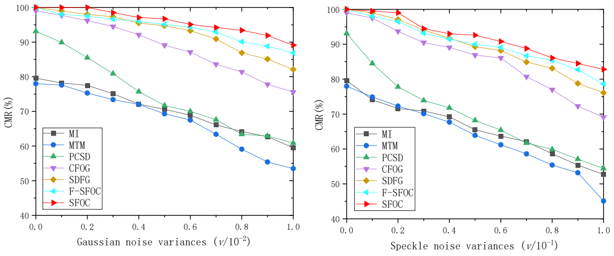


Figure 6. Matching results of all test cases by SFOC with the template size of 80 × 80 pixels. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4. (e) Case 5. (f) Case 6. (g) Case 7. (h) Case 8.

different levels of Gaussian white noise and speckle noise to the images, respectively. Because the NRD between multimodal images is difficult to be precisely fitted only by a simple mathematical model. Meanwhile, LiDAR and SAR images typically contain more noise than infrared images, which is not conducive to the assessment of noise sensitivity. Consequently, all the methods were performed with the template size of 80 × 80 pixels for the selected four pairs of Optical-to-Infrared cases, and their average value of CMR was used for the subsequent analysis. Specifically, two types of series noisy images were generated by adding the different levels of Gaussian white noise with mean 0 and variance v in the range [0, 0.01] with an interval of 0.001, and the different levels of speckle-noise with variances v in the range [0, 0.1] with an interval of 0.01, respectively.

Figure 7 presents the average CMRs of different methods versus various noise consisting of Gaussian white noise and speckle noise. SFOC and its degraded version (i.e., F-SFOC) achieved superior capacities under increasing Gaussian and speckle noise, followed by SDFG and CFOG. It demonstrated that the generation of SFOC using the dilated Gaussian convolution with different dilated rates could be more effective for resisting noise than SDFG only utilizing the general Gaussian convolution, and the generation of SFOC and SDFG both using a series of filters was more useful in withstanding noise than CFOG only utilizing simple gradient computation with the pixel difference. While the orientation channels of CFOG were implemented by the Gaussian kernel, which is more effective to reduce the interference of noise than PCSD. In addition, the performance of MI was relatively stable under various noises, but its average CMR was still lower than SFOC, F-SFOC, and CFOG. And MTM also presented lower robustness to Gaussian and speckle noise compared with MI.



(a) (b)

Figure 7. Average CMRs of different methods versus various noise. (a) Average CMRs of different methods versus various Gaussian white noise. (b) Average CMRs of different methods versus various speckle noise.

The above results and coherence analysis demonstrate that SFOC has apparent effectiveness and advantages for resisting significant NRD and noise between multimodal images, as well as high computational efficiency. The good adaptive performance was mainly due to the following reasons. On the one hand, it not only employed the first-order steerable filters with the multi-scale strategy to depict the multi-directional and multi-scale structural features between multimodal images, but it also utilized the second-order steerable filters and three parallel dilated Gaussian kernels to emphasize more detailed information and multi-level context structural features, respectively, which further improves the discriminative and anti-noise capability of the proposed method. On the other hand, the improved Fast-NCC$_{SFOC}$ based on the FFT and integral image technique ensured its fast computational efficiency.

4. CONCLUSIONS

This paper presented a robust matching method of multimodal RS images, involving both a novel SFOC descriptor and a fast similarity measure (i.e., Fast-NCC$_{SFOC}$). SFOC is first proposed by making use of the first- and second-order Gaussian steerable filters, which aims to capture distinctive multi-directional and multi-scale structural features for resisting significant NRD between multimodal images. Then Fast-NCC$_{SFOC}$ is established by combining NCC and SFOC, and it speeds up the image matching process by using the FFT technique and the integral image. The experimental resulted on eight various multimodal images have demonstrated the robustness and effectiveness of SFOG. In contrast to other state-of-the-art methods (i.e., MI, MTM, PCSD, CFOG, SDFG), the proposed SFOC achieved the best matching performance in the quantitative evaluation.

Although SFOC presented robust performance for multimodal image matching, it is sensitive to global geometric distortions between images, that is, it cannot be adapted to multimodal image matching with large scale or rotation differences. Future research aims to design an enhanced descriptor that is adaptable to geometric distortions without the assist of geo-referenced information, and explore the matching technique with scale and rotation invariance.

ACKNOWLEDGEMENTS

This paper is supported by the National Natural Science Foundation of China (No.41971281).

REFERENCES

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834-848.

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., Tupin, F., 2014. SAR-SIFT: a SIFT-like algorithm for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(1): 453-466.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A. and Sattler, T., 2019. D2-net: A trainable CNN for joint description and detection of local features. *In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 8092-8101.

Fan, J., Wu, Y., Li, M., Liang, W., Cao, Y., 2018. SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9): 5368-5379.

Fan, Z., Zhang, L., Liu, Y., Wang, Q., Zlatanova, S., 2021. Exploiting High Geopositioning Accuracy of SAR Data to Obtain Accurate Geometric Orientation of Optical Satellite Images. *Remote Sensing*, 13(17): 3535.

Freeman, W.T., Adelson, E.H., 1991. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9): 891-906.

Hel-Or, Y., Hel-Or, H., David, E., 2013. Matching by tone mapping: Photometric invariant template matching. *IEEE transactions on pattern analysis and machine intelligence*, 36(2): 317-330.

Huang, D., Zhu, C., Wang, Y., Chen, L., 2014. HSOG: a novel local image descriptor based on histograms of the second-order gradients. *IEEE Transactions on Image Processing*, 23(11): 4680-4695.

Jiang, Y.H., Zhang, G., Chen, P., Li, D.R., Tang, X.M., Huang, W.C., 2015. Systematic error compensation based on a rational function model for Ziyuan1-02C. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7), pp.3985-3995.

Li, J., Hu, Q., Ai, M., 2019. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29: 3296-3310.

Lowe, D.G., 2004. Distinctive image features from scale-invariant key points. *International journal of computer vision*, 60(2): 91-110.

Ma, J., Zhou, H., Zhao, J., Gao, Y., Jiang, J., Tian, J., 2015. Robust feature matching for remote sensing image registration via locally linear transforming. *IEEE Transactions on Geoscience and Remote Sensing*, 53(12), 6469-6481.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. *In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*. IEEE, pp. I-I.

Wallis, S.A., Georgeson, M.A., 2012. Mach bands and multiscale models of spatial vision: the role of first, second, and third derivative operators in encoding bars and edges. *Journal of vision*, 12(13): 18-18.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G., 2018a. Understanding convolution for semantic segmentation. *In: 2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 1451-1460.

Wang, S., Quan, D., Liang, X., Ning, M., Guo, Y., Jiao, L., 2018b. A deep learning framework for remote sensing image registration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145, pp.148-164.

Ye, Y., Bruzzone, L., Shan, J., Bovolo, F., Zhu, Q., 2019. Fast and robust matching for multimodal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11): 9059-9070.

Ye, Y., Shan, J., Bruzzone, L., Shen, L., 2017. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5): 2941-2958.

Zhou, L., Ye, Y., Tang, T., Nan, K., Qin, Y., 2021. Robust Matching for SAR and Optical Images Using Multiscale Convolutional Gradient Features. *IEEE Geoscience and Remote Sensing Letters* (Early Access), 1-5.

Zhu, B., Ye, Y., Zhou, L., Li, Z., Yin, G., 2021. Robust registration of aerial images and LiDAR data using spatial constraints and Gabor structural features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181, pp.129-147.